ARI Research Note 99-30

Documentation and Archival of Selected ARI Data Bases Phase II: Final Summary Report

Ani S. DiFazio and Winnie Y. Young
Human Resources Research Organization

Dianne P. Driessen and Donna Peck
Fu Associates, Ltd.

Organization and Personnel Resources Research Unit
Paul A. Gade, Chief

June 1999

19990701 047

U.S. Army Research Institute
for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

DTIC QUALITY INSPECTED 4

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE June 1999	3. REPORT TYPE AND DATES COVERED Final Report, April 2, 1997 - August 30, 1998

4. TITLE AND SUBTITLE
Documentation and Archival of Selected ARI Data Bases, Phase II:
Final Summary Report

5. FUNDING NUMBERS
MDA903-93-D-0032 DO 0057
6900C06/A792 0603007A

6. AUTHOR(S)
Ani S. DiFazio and Winnie Y. Young (HumRRO); Dianne P. Driessen
and Donna Peck (Fu Associates, Ltd.)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

HumRRO
66 Canal Center Plaza, Suite 400
Alexandria, Virginia 22314

8. PERFORMING ORGANIZATION REPORT NUMBER

FR-EADD-98-53

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

Research Note 99-30

11. SUPPLEMENTARY NOTES

Contracting Officer's Technical Representative, Dr. Ronald B. Tiggle

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited.

12b. DISTRIBUTION CODE

13. **ABSTRACT** ***(Maximum 200 words)***

Since 1975, the U.S. Army Research Institute (ARI) has collected a wide array of Manpower Personnel Research (MPR) and Training data in support of its research activities. Until this current effort, there were no formal procedures or guidelines for documenting and archiving these numerous databases. The ability of new users to access and use extant ARI data was heavily dependent on the knowledge of those ARI staff members who worked most closely with the data. With organization turnover and downsizing, critical information needed to access and use data would have been lost over time. This project had two phases. The first Phase developed documentation and archive standards for extant and future ARI data. Phase II, the focus of this report, applied those standards to specified extant data. The technical approach and procedures used in Phase II of this project is the subject of this report.

14. SUBJECT TERMS
dataset, data file, database, documentation, codebook, user's guide, archive

15. NUMBER OF PAGES
70

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL

FOREWORD

Since 1975, the Army Research Institute (ARI) has developed numerous databases in the course of conducting research. Until this effort, there were no formal guidelines for the documentation and archive of these data. Due to the lack of such guidelines, many databases were neither documented nor archived at a central location. As a result, the ability of new users to access and utilize these data was heavily dependent on the knowledge of those ARI staff members most familiar with the data. With organizational turnover and downsizing, the information needed by new users to access and use ARI data was being lost over time. Clearly, documentation and archive guidelines are necessary for future ARI databases. In addition, in order to preserve valuable extant ARI data, it was necessary to apply these documentation and archive standards to existing databases.

This report provides a project summary of the second of a two phase project designed to document and archive ARI databases. The first phase focused on the development of documentation and archive standards. This second phase focused on the application of those standards to extant ARI databases. This report provides a detailed description of the technical approach used in the documentation and archive of existing data. In addition, it provides a detailed description of the hard copy and electronic deliverables associated with this project.

Each task originally planned for this project has been completed.

ZITA M. SIMUTIS
Technical Director

ACKNOWLEDGMENTS

The authors acknowledge and appreciate the helpful cooperation of Dr. Ronald Tiggle, the Technical Monitor of this Delivery Order, and all the ARI staff who served as dataset Points of Contact.

DOCUMENTATION AND ARCHIVE OF SELECTED ARI DATA BASES, PHASE II: FINAL PROJECT SUMMARY REPORT

EXECUTIVE SUMMARY

Requirement:

The Army Research Institute (ARI) is committed to the preservation of the numerous databases that have been created since 1975 in support of its research activities. Prior to this effort, there were no formal guidelines or standards for the documentation and archive of these data. New users interested in accessing extant data had to rely in large part on the knowledge of ARI staff for necessary information. If this effort had not been undertaken, the information needed by new potential users of ARI databases would be lost over time because of organizational turnover and downsizing. In some cases, the actual data would have been lost as well. Clearly, documentation and archive standards are necessary to describe and preserve ARI data. The development of these standards was the focus of Phase I of this effort. Phase II, the focus of this report, implemented the documentation and archive standards developed during Phase I.

Procedure:

A hierarchical system of three documentation levels was developed in Phase I based on review of the literature and ARI needs. This documentation system is designed to be sensitive to the trade-off between resources required in generating higher levels of documentation and the utility of that documentation to future users of ARI data. The lowest level or baseline documentation was completed for all designated Manpower Personnel Research (MPR) and Training datasets. This information is stored in a Microsoft Access database and contains basic descriptive information for datasets. A stand-alone version of the database was also produced to facilitate key entry of baseline information for future datasets. The second level of documentation involved the organization of existing hard copy materials for extant datasets. The third level of documentation consists of the development of detailed codebooks for selected ARI data files.

With the exception of some datasets specified by ARI, all data files were archived on CD-ROM. A CD directory structure and README file standard were adopted. All archived files underwent extensive and rigorous quality assurance checks. Seventy-two ARI projects are represented among the data archived in this effort. Baseline documentation has been completed for 143 datasets and their 542 component data files. Of the 143 datasets, 137 datasets were identified for archive; from them, 354 individual data and format files have been archived on CD-ROM.

Utilization:

There is now an easily accessible cache of baseline information on all identified MPR and Training datasets. The design of the system allows for the easy integration of baseline information of future datasets. Hard copy materials for selected datasets have been organized and archived. Codebooks have been produced for selected data files following the documentation standards developed in Phase I. In addition, there is now a data archive of ARI files that is centralized, informative, uniform, and easy to use.

TABLE OF CONTENTS

List of Tables

Page

List of Appendices

DOCUMENTATION AND ARCHIVE OF SELECTED ARI DATA BASES, PHASE II: FINAL SUMMARY REPORT

Introduction

Since 1975, the U.S. Army Research Institute (ARI) has collected a wide array of Manpower Personnel Research (MPR) and Training data in support of its research activities. Until this current effort, there were no formal procedures or guidelines for documenting and preserving these numerous databases. As a result, many extant datasets were neither documented nor archived in a central location. The ability of new users to access and utilize extant ARI data, whether collected by ARI staff or by outside contractors, was heavily dependent on the knowledge of those ARI staff members who had worked most closely with the data. With organizational turnover and downsizing, critical information needed to access and use data would have been lost over time had this current effort not been undertaken.

Objectives

The goal of this project was to preserve and document datasets, so that future researchers can use the data collected by or for ARI efficiently and effectively. This effort had two phases. The development of documentation and archive standards for extant and future ARI datasets was completed during Phase I. Phase II, the focus of this report, implemented the standards established in Phase I for the documentation and preservation of extant data.

Technical Approach

The technical approach of this project was formulated during Phase I and is described in our report *Final Project Summary Report – Phase I* (DiFazio, Young, & Driessen, 1997a). The present report describes the steps taken in Phase II for the practical application of the technical approach developed in Phase I. First, however, two concepts that are central to our technical approach are discussed: documentation level and data entities.

Documentation Level

One objective of Phase I was to produce a system of documentation standards whereby the appropriate level of documentation would be generated for each specified extant dataset. Given limited resources, it was clear that the maximum level of documentation could not be justified for all extant data files. Some data files are deemed important by ARI and are likely to be used again, while others are not. Some files may have an abundance of supporting technical documentation, descriptive variable names, labels, and formatted values, which make them easy to understand and use. Other data files with uninformative variable names (e.g., V1, V2) and/or no descriptive labels, format values, or supporting documentation will probably not be useful to a new user.

The documentation system that we developed had to be responsive to the varying condition and status of extant data files.

Based on discussions with ARI staff, our knowledge of ARI data files, and a review of the literature, we formed a three-level system of documentation. This system was designed to be hierarchical, in that all files would be documented at the highest assigned level, as well as all lower levels. Documentation levels are discussed in detail in DiFazio and Young, (1997b). Table 1 summarizes the three levels of documentation.

Table 1. Documentation Levels

Levels	Level I	Level II	Level III
I Conduct Baseline Documentation	Yes	Yes	Yes
II Archive Hard Copy Materials	No	Yes	Yes
III Generate Codebook	No	No	Yes

The Data Documentation section below describes the characteristics of each documentation level. The final assignment of documentation level to datasets was made by ARI. Also, due to budgetary limitations, ARI specified which datasets would be documented as part of this effort. The final list of datasets and their documentation level assignments are presented in Appendix A. It should be noted that the concept of documentation level was pivotal in guiding the documentation activities undertaken during this project dealing with extant ARI data. Its usefulness does not extend beyond the duration of this project, since the recently promulgated ARI Regulation 70-31 dictates a uniform approach to the documentation and preservation of all future ARI data files.

Data Entities

Some mention of the terminology used in this report is desirable, as general usage varies. The term "dataset" is commonly used in the industry to mean a single physical electronic file and is often used interchangeably with the term "data file." The term "database" is typically used to refer to a family of related data files. For example, the Project A database comprises numerous individual data files. Data files within databases may or may not be structured differently from one another; typically, they can all be merged together by one or more "link" variables.

At the beginning of this project, we found that, in addition to a single physical file, ARI researchers used the term "dataset" to also refer to a collection of similar, conceptually grouped, albeit physically distinct files. For example, a set of longitudinal files where only data collection year differentiates the physical files is typically termed a "dataset" by ARI staff; the components of this "dataset" are referred to as data files. Thus, each year of data in the Enlisted Master File (EMF) is referred to as a "file" and the collection of EMF "files" as a "dataset." Strictly speaking, the EMF is a database and the yearly data, datasets or data files. For clear communication, the terms "dataset" and "data file" will be defined as ARI researchers have defined them throughout the remainder of this report.

It is important to note that the different documentation levels used in this project describe different data entities. While baseline documentation and the organization of available hard copy materials (documentation levels I and II) describe datasets, codebooks (documentation level III) describe data files. Also, the smallest data entity that can be archived is a data file.

Data Documentation

Level I or Baseline Documentation

Objective. Baseline or level I documentation is the minimum level of documentation for all specified extant MPR and Training datasets. The purpose of baseline documentation is to provide on-line technical information for all datasets and allow researchers the opportunity to gain basic dataset information quickly and efficiently. The type of information collected as part of baseline documentation includes:

- Available Documentation
- Information on Parent Database
- Dataset Access Issues
- Data Description
- Technical Dataset Information
- Project Research and Analyses
- Publications
- Technical Information on Component Data Files

The ARI Research Dataset Questionnaire. Baseline documentation information for extant datasets was collected from dataset Points of Contact (POCs) through an in-depth survey called the ARI Research Dataset Questionnaire. The questionnaire, originally developed during Phase I, was modified during Phase II in several ways. First, since the survey was initially designed for MPR datasets, it was expanded to include information pertinent to Training data. Second, items that had been included specifically to assist project staff in archiving extant data files and to facilitate the assignment of documentation levels to datasets were eliminated. Third, unlike Phase I, when the questionnaire was administered in person to dataset POCs by the project data collection team, Phase II budgetary considerations dictated that the survey be self-administered, with advance and follow-up contact by the data collection team. Therefore, some survey items and directions were reworded in Phase II so that dataset POCs could complete the questionnaire without the presence of project staff.

Once baseline data had been collected on all specified extant datasets, the survey was modified to support ARI's on-going need to document and archive future datasets. Survey adjustments were made to simplify the entry process, ensure consistency with existing information, and accommodate the differences in the technical environments of past and future automated data. For example, while many of the extant data files archived in this project were resident on the National Institutes of Health (NIH) computer facility at the onset of this project, Regulation 70-31 governing future ARI data stipulates that new data files will be submitted for archive already on CD-ROM. Therefore, we eliminated unnecessary questions and options associated with mainframe processing and formats. In spite of these adjustments, changes were kept at a minimum to control cost and ensure consistency of information. The final questionnaire is presented in Appendix B.

Information Collection. An initial list of 148 MPR and 56 Training datasets were identified during Phase I. Of these 148 MPR datasets, baseline documentation was completed for 38 datasets during Phase I.

The project team began Phase II by contacting dataset POCs of the remaining 166 MPR and Training datasets to discuss the data collection process. Each POC at ARI Headquarters then received an ARI Research Dataset Questionnaire for self-administration. The Phase II data collection process was iterative in nature. POCs completed the surveys and returned them to the data collection team. Once they had reviewed survey responses, the data collection team made several calls or visits to each POC to ask questions, fill in missing or inconsistent information, and pick up hard copy materials and data, if available. Many POCs had to look up information, consult with research team members or technical staff, make copies of data files, and locate documentation. While some POCs handed data and materials directly to the data collection team, others dropped these off in a secure room at ARI designated as a holding area for data and materials. The data collection team routinely picked up the materials/data deposited in this holding room. They maintained a detailed log of the hard copy materials and data files received from POCs for each dataset.

Baseline documentation for training data from ARI at Fort Rucker was collected in some cases by the self-administered survey and in other cases through personal interviews with dataset POCs. Baseline documentation from all other ARI locations was collected through mailed self-administered surveys with follow-up telephone calls from the data collection team. In these cases, data and hard copy materials were received from the POCs by regular or electronic mail.

As the project team members worked with dataset POCs during Phase II to continue with baseline documentation, the list of MPR and Training datasets underwent modification. With the assistance of the project's Contracting Officer's Technical Representative (COTR), as the project unfolded some datasets were added and others deleted from the MPR and Training list, resulting in a final total of 143 datasets. The ARI Research Documentation Database contains baseline information on all of these datasets.

The ARI Research Documentation Database. Information for baseline documentation is stored in a Microsoft Access database called the ARI Research Documentation Database. The database was designed during Phase I and modified in Phase II. In addition to changes necessitated by the alterations to the ARI Research Dataset Questionnaire described above, the ARI Research Documentation Database was redesigned during Phase II to include:

- Updated README information that includes survey examples, definitions of terms, navigation tips, and a diagram describing data entities or units of analysis,

- A date field to be updated by the ARI archivist to reflect the recentness of the recorded information,

- Redesigned and consolidated forms and reports to improve data presentation, instructions, and navigation,

- Integration of a tables/form/report into the database to record available documentation for inventory control.

After the data collection process was completed, all information from the surveys was entered into the ARI Research Documentation Database. Some information from the dataset POCs had to be reviewed and assimilated before key entry. To reduce error and promote efficiency, special Microsoft Access forms were created to make the key entry screens match the survey format as closely as possible. Once the ARI Research Dataset Questionnaire responses were key entered, the automated data were then reviewed by project analysts to assure quality control. Upon completion of this quality review, the information contained in the database was printed and sent to the dataset POCs for review, verification, and comment. All POC corrections resulting from this review were made to the database.

Additional quality assurance measures were undertaken . The information that the dataset POCs provided reflected the characteristics of the data as they were known prior to archive. While this information was very helpful in our archive efforts, it did not accurately reflect the characteristics of the vast majority of files once they were archived. A detailed discussion of this and other archive issues is provided below. Since one of our objectives was to have the ARI Research Database reflect the status and characteristics of archived data, we modified the database to provide updated archived data file information.

The ARI Research Documentation Database underwent a series of tests designed to identify and correct programming errors. The key entry process and the production of reports reflecting the survey data and materials collected for each dataset provided a first tier of testing of the design and programming of the database. Project programming staff then began explicitly testing the functionality and operation of navigation buttons, entry screens, forms, and reports in the first or alpha version of the database to ensure programming accuracy. This extensive process provided a thorough testbed for identifying programming problems and making improvements in key entry. After the necessary adjustments were made, the COTR and other project staff were given the software and database for review and comment. The minor modifications suggested as a result of this review were implemented and a final status version of the ARI Research Documentation Database was produced.

Once the software and database were finalized, the Administrative ARI Research Documentation Database was produced. The Administrative Database includes the final database with security features that restricts write-access to the ARI Archivist; the security features provide other ARI researchers with read-only access to the ARI Research Documentation Database.

In addition, a stand-alone version of the software was produced that allows future researchers to directly enter baseline dataset information using Microsoft Access forms. The information entered in the stand-alone version will produce a single dataset "record" that can be appended to the full database by the ARI Archivist using procedures outlined in documentation accompanying the Administrative Database. This set of software allows the ARI Archivist full and secure control over the contents of the Administrative Database, but shifts the baseline information key-entry function from the Archivist to the researcher most familiar with the dataset.

As stated above, entries in the ARI Research Documentation Database are stored at the dataset level, although unique data file information is presented for each component data file in a dataset.

Level II or Hard Copy Documentation

Objective. Level II documentation is a collection of the hard copy extant materials produced as part of the original research project involving the dataset. The purpose of Level II documentation is to provide supplemental information that is too lengthy or complex to be included in baseline documentation. The types of hard copy documentation collected include materials describing:

- Research Design
- Sampling Plan and Procedures
- Data Collection Plan
- Data Collection Instruments
- Data Development Process
- Codebook or User's Guide
- Record Layout(s)
- Research Report(s)/Publication(s)

As discussed previously, ARI assigned the documentation level for datasets and their component data files. As Appendix A indicates, of the 28 datasets assigned a level II-only documentation status, 23 had POC-supplied hard copy documentation for which Level II documentation could be completed.[1] In addition, since documentation levels are

[1] The five datasets assigned a Level II-only status that had no POC-supplied materials are #53.00 (Gender Integrated Basic Combat Training), #57.13 (Multinational Force Observers –Sinai: Job Knowledge Dataset #2), #68.00 (Special Forces SFAS Longitudinal Validation), #83.01 (Bosnia Pre-Deployment Soldiers Survey), and #530.00 (A Prototype Army National Guard (ARNG) Armor and Mechanized Infantry Gunnery Training Assessment Database).

hierarchical, the 14 selected Level III files were Level II documented as well. Therefore, Level II documentation was produced for 37 datasets.

Collection of Materials. Level II documentation was collected at the dataset level. In most cases, dataset POCs provided the actual materials that they reported as existing for a particular dataset in the ARI Research Dataset Questionnaire with the survey itself. In some cases, the project data collection team prompted the POCs for those materials. For ARI headquarters, the documents were either handed directly to project data collection staff by the POC or dropped off in the ARI holding area described previously. Materials from other ARI locations were sent to the project data collection staff via regular mail or other mail handling service. A detailed log was maintained of the hard copy materials received from POCs for each dataset.

Organization of Materials. A goal of Level II documentation was not just the collection but also the organization of hard copy materials for extant datasets. We took a number of steps to achieve this end. First, all documents were labeled with the dataset number so that the dataset attribution of the document would never be in question.[2] Second, each document submitted by a dataset POC was reviewed and organized by a project staff member trained in Library and Information Science. In its organized form, the first hard copy document for each dataset is a report called the ARI Research Dataset Summary, which contains hard copy baseline information for the dataset from the ARI Research Documentation Database.

Next, hard copy documents were sequenced to follow the research process in general. For example, documents pertaining to research and sampling design and data collection instruments were placed before the final report. For most datasets, the last document(s) is a hard copy listing of a SAS PROC CONTENTS and/or SPSS SYSFILE INFO for data files associated with the dataset.[3] These listings were generated by those project staff responsible for data file archive. Once materials were organized in this fashion, indices of hard copy documents were generated for each dataset. Indices contain the dataset number, name, and a numbered bibliographic list of hard copy documents. A sample index is provided in Appendix C.

Depending on the length of materials, documents were either bound by GBC plastic spiral binding or placed within folders or computer binders,. A dataset may comprise a number of these bound sets of documents or binders. For ease of reference, a label identifying the dataset number, the binder number, the total number of binders for that dataset, and the index item number(s) contained in that binder was affixed to the upper right corner of each binder.

[2] In addition to other materials, we were given seven file folders containing miscellaneous documents such as telephone messages, e-mail and regular correspondence, printout, and notes for dataset # 42.00 (The 1990 Army Career Satisfaction Survey). In order to dataset-identify these materials, we bound the documents within each of the seven folders together and affixed a dataset number label on the top page.

[3] This is true for component files of all datasets except Level I flat files and dataset #503.00 (National Training Center Archive), which consists of thousands of files such as FoxPro, Microsoft Word, and dBase.

As stated previously, those datasets assigned a documentation Level II or III (see Appendix A) were originally slated as the only datasets for which hard copy documents would be collected. As our information collection efforts for datasets of all documentation levels began, however, it became clear that many POCs of Level I datasets wanted to submit their hard copy documentation for centralized archive. Hard copy materials were submitted for all of the 69 Level I datasets (see Appendix A). After discussing this unforeseen turn of events with the project COTR, we created a category of hard copy documentation processing called a "level I pass-through." Like Level II and III datasets, we maintained a log of Level I dataset hard copy documentation, affixed the dataset number on each document, and produced a PROC CONTENTS and/or SYSFILE INFO listing on component files, when appropriate. Unlike Level II and III datasets, the Level I pass-through documents were then simply placed in manila envelopes and labels were affixed to the upper left corner containing dataset number, name, the manila envelope number, and the total number of envelopes for that dataset.

Level III or Codebook Documentation

Objective. Level III documentation, the highest level of documentation, consists of a codebook or user's guide that contains data file as well as variable descriptions and information. Three key components of Level III documentation are:

- File Descriptions and Data Element List
- Data Element Descriptions
- Format Descriptions, where applicable

As noted previously, ARI assigned documentation level for datasets and their component data files. As Appendix A indicates, of the 46 datasets assigned a Level III documentation status, 14[4] were selected for codebook development in this project[5]. All Level III documentation is at the data file level. The 46 datasets assigned Level III status were comprised of 48 component data files; the 14 selected datasets were comprised of one data file each.

Codebook Development. Codebooks were developed for SAS, SPSS, and flat files.[6] For SAS and SPSS files, the components of the codebooks were generated in the language in which the data were stored. Codebooks for flat files were generated using SAS. A special program was written to read both SAS and SPSS output and automate the production of tables of contents to ensure accurate page number references. These components of codebooks were then pulled into Microsoft Word for assimilation and

[4] In addition to these, there were 3 datasets composed of 8 data files with pre-exiting ARI-produced codebooks. These datasets are #32.00 (Longitudinal Research on Officer Careers), #65.00 (survey of Officer Careers), and #100.00 (Nelson-Denny Reading Test Scores – West Point, U.S.M.A). Also, although Level I, electronic codebooks were available for all datasets for Project #20 (Project A/Building the Career Force).

[5] The hard copy materials submitted by the dataset POCs for those Level III datasets not selected for Level II documentation in this effort were treated as Level I pass-throughs.

[6] Since we were producing codebooks in SAS for flat files, we included the SAS transport version of the flat file, as well as the SAS creation program, in the archive.

cover pages were generated. Each codebook was stored as a Microsoft Word document in two volumes.[7] The first volume of a Level III codebook consists of general information on the datafile and variables; the second provides descriptive statistics:

- Volume 1 (CBK1.DOC)
  - Cover Page
  - Table of Contents for SYSFILE INFO listing for SPSS Data Files
  - SAS PROC CONTENTS or SPSS SYSFILE INFO

- Volume 2 (CBK2.DOC)
  - Cover Page
  - Table of Contents for Variable Descriptive Statistics
  - Variable Descriptive Statistics and Format Values

Data Preservation

In addition to data description or documentation, the preservation of datasets was one of the primary objectives of this project.

Data File Collection. Dataset POCs were asked to list the data files that were developed for each identified dataset in the ARI Research Dataset Questionnaire. The 143 datasets contained in the ARI Research Documentation Database come from 72 ARI projects. Of these 143 datasets, 137 had files transmitted to us during the course of this project.[8] POCs with datafiles in their possession were then asked to transmit the data to the project data collection team directly or deposit the file(s) in a secure room at ARI headquarters designated as a holding area. Detailed logs of data files received from POCs by the project data collection team were maintained. During the course of this project, 125 files recorded on microcomputer media were supplied to us by POCs. An additional four files were transmitted via electronic mail.

Project staff were given direct access to the datasets resident on the NIH computer through the use of an account at the facility. The information given by the POCs in response to the ARI Research Dataset Questionnaire proved invaluable in facilitating our access of these data files. There were 225 files either resident on the NIH computer or

[7] For those three datasets with full ARI-produced hard copy codebooks, a CBK1 file was included in the archive as a courtesy.

[8] Datasets # 3.01 (Special Forces Assessment and Selection Database), #3.02 (Special Forces Qualification Course Class Database), and #53.00 (Gender Integrated Basic Combat Training) are close-hold datasets and, therefore, not a part of the ARI Research Dataset Archive. Also, datasets #83.02 and #83.03 (part of the Operation Joint Endeavor Project) were not completed as of the end of this project and, therefore, are not currently part of the archive. In addition, ARI determined that datasets #27.00 (EMF Extract), #28.00 (Cohort Data Base), and #29.00 (Applicants and Accessions File) would be archived by in-house ARI staff. Therefore, while all these datasets appear in the ARI Research Documentation Database, they are not currently part of the ARI data archives.

recorded on mainframe media. Therefore, the total number of component files to the 137 datasets for which we received data files is 354.[9]

Recording Medium. A detailed review of the literature regarding data recording media is presented in DiFazio and Young (1997b). Based on this review of the literature and an understanding of ARI's needs, the CD-ROM[10] was designated as the archive data recording medium for the extant data archived in this project as well as future ARI data files. The CD-ROM used to record extant data were formatted in ISO 9660 format and the DOS eight character name and three character suffix file naming convention was used.

Mainframe Media File Preparation. All data supplied directly by POCs that were recorded on mainframe media were on 3480 cartridge tapes or 9-track tapes. These files were taken to the NIH computer facility and logged into the tape library. Those ARI files already resident at NIH and slated for archive were also recorded either on tape media and were, therefore, part of the NIH tape library, or on disk. Of these files that required some mainframe processing, there were 103 flat, 116 SAS[11], and 6 SPSS files. The files were then accessed to make sure that the files were not corrupted and that the data were readable.

ARI requested that files with encrypted unique identifiers, such as Social Security Number (SSN), be de-encrypted before archive whenever possible. There were 45[12] mainframe data files with encrypted unique identifiers for which encryption algorithms were available. Some had originally been encrypted by code written in PL-1, while others had been encrypted using a SAS program. We converted the PL-1 program into SAS code, tested this conversion for accuracy, and then used the SAS program to de-encrypt the unique identifiers. For those files originally encrypted by a SAS program, we simply used that program to de-encrypt the identifiers.

All of these 122 SAS and SPSS system files had been originally recorded in non-transport mode. Since one of the goals of data archive was to ensure the usability of data across host computer environments and operating systems, we created copies of SAS and SPSS system files that were written in transport or portable mode. Essentially, this procedure converts system files into sequential files. Data originally written as flat files were not modified. Once we had transport versions of SAS and SPSS system files, the data were ready to be downloaded.

[9] The 354 component files include format libraries as well as data files. This figure does not include the thousands of files, including Foxpro, dBase, and Microsoft Word files, archived for the National Training Center dataset #503.00.

[10] RICOH's CD-R Type 74 with 650 MB was selected as the specific CD-ROM on which to archive extant data.

[11] This includes both SAS data files as well as format libraries.

[12] These files were associated with datasets #8.01 (ROTC Commissioned Dataset), #8.02 (U.S. Military Academy Dataset), #8.03 (ROTC Advanced Camp Dataset), #8.05 (Automated Instructional Management System Dataset), #8.06 (Officer Core Dataset), and all 5 datasets for Project 20 (Project A/Building the Career Force).

The process of downloading mainframe data across commercial telephone lines began using a widely used file transfer protocol called KERMIT. Due to the security limitations of public (non-dedicated) telephone lines, however, ARI asked that we stop downloading data in this manner shortly after we began. We investigated a number of alternatives for transferring these data, including the use of hard-wired mainframe to microcomputer connections at NIH. Finally, we decided to use the services of a conversion facility. Essentially, this facility has the capability of transferring data from mainframe tape media to CD-ROM without going through commercial telephone lines. This method of data transfer is safe, efficient, and accurate. As an added precaution, each file that the conversion service wrote onto CD-ROM was checked for usability and accuracy.[13]

The IBM file naming conventions used by the NIH computer facility are vastly different from the DOS eight character name plus three character suffix file nomenclature adapted in this project for archived data. The external name of an NIH file begins with the registered account and initials of the user. The remainder of the name consists of two or more groups of one to eight characters, separated by periods. The entire file name cannot exceed 44 characters. In addition to external dataset names, NIH files also contain internal members with names of up to eight characters each. Project analysts/archivists chose the most descriptive part of the external and internal NIH dataset name in arriving at the archive file name.[14] For SAS files, the suffix given the archived files was always XPT to denote transport file. Archived SPSS files have the suffix POR to denote portable files. Flat files downloaded from the mainframe always contain the suffix ASC.

Microcomputer Media File Preparation. We received 11 flat, 13 SAS, 96 SPSS, and 5 EXCEL files from POCs recorded on microcomputer media. As on its mainframe counterpart, data recorded on microcomputer media were initially accessed to assess usability.[15] There were 3[16] data files presented to us on microcomputer media for which an encryption algorithm for unique identifiers was available. The SAS program based on the PL-1 code used to de-encrypt unique identifiers on mainframe files was used to de-encrypt unique identifiers on microcomputer data.

SAS and SPSS files were then copied and written in transport mode. Data originally written as flat files were not modified. Like SAS and SPSS data files with mainframe origins, we assigned suffixes of XPT and POR to denote portability of archived files originally created and stored on the microcomputer. The original names, including the suffix, of flat files were retained.

[13] Each data file converted from mainframe tape media to CD-ROM was verified for accuracy. SAS transport files were converted to SAS system files and PROC CONTENTS were generated and compared to that produced on the mainframe. SPSS transport files and SAS format libraries were verified in the same manner.

[14] For example, an original NIH data file name was WRZ1KFD.OMF94.SAS.OFFICER with an internal member name of OMF. We recorded a transport version of this file on CD-ROM called OMF94OF.XPT.

[15] Only one file, the format library for dataset # 5.02 (STAMP), was discovered to be damaged internally and unreadable.

[16] These files were associated with datasets #6.01 (Enlisted Core Dataset - 100%) , #6.02 (Enlisted Core Dataset – 25%), and #8.04 (Officer Longitudinal Dataset).

CD Directory Structure. The COTR determined that the data archives should be organized at the ARI project level. A list of ARI projects represented by datasets included in this effort is contained in Appendix D. Datasets were numbered so that the integer portion of the number represents the project number. For example, datasets #87.01 and #87.02 are part of project #87 Leadership Requirements Analysis. Since each project can comprise a number of datasets and each dataset can contain a number of data files, we created a CD-ROM directory structure that allows the user to quickly identify the data file(s) of interest.

The volume label of the root directory for each CD contains the project number following the word "PROJECT_". For example, the volume label for the CD-ROM containing data for the Leadership Requirements Analysis is "PROJECT_087". Each dataset within a project has its own directory. These dataset directories are named "DS" followed by the dataset number, where the decimal point of the dataset number is replaced with an underscore. For example, dataset 87.01 is contained in a directory called DS087_01 on the CD-ROM with the volume label PROJECT_087. These "DS" directories contain subdirectories for all associated data files. These data file subdirectories are named "DF" and are numbered sequentially within the "DS" directory, starting with 001.[17] The following paradigm depicts a project consisting of three datasets, where the first dataset contains two data files, the second three files, and the third one data file.

```
Project_xxx (CD Root Directory)
      |
      |→DSxxx_xx
      |       |
      |       |__|→DF001
      |          |→DF002
      |
      |→DSxxx_xx
      |       |
      |       |__|→DF001
      |          |→DF002
      |          |→DF003
      |
      |→DSxxx_xx
              |
              |__|→DF001
```

The CD root directory contains only "DS" directories. In addition to "DF" subdirectories, "DS" directories contain a dataset-level README file. "DF" subdirectories contain data file-level README information, data, and formats, codebooks, and other files, where available.

[17] In addition to data files, SAS format libraries were given their own "DF" subdirectories.

README Files. ASCII README files were created to provide easy access information to the user. As stated above. README files were created for both datasets ("DS" directories) and data files ("DF" subdirectories). The "DS" README contains the following information:

- Introductory paragraph
- Dataset number
- Dataset description
- Project/analysis name[18]
- Archive date
- Number of data files associated with dataset
- Names and contents of files and subdirectories
- Information on which datafiles a SAS format library is associated with, where applicable
- Information on multiple CD datasets, where applicable

"DF" README files contain the following information:

- Introductory paragraph
- Dataset number
- Project/analysis name
- Archive date
- Data file name
- Contents of subdirectory, including:
  - Archive data file name
  - Original data file name, where different from archive name
- If SAS data, sample code to convert transport file to system file
- If SAS data with a format library, sample code to convert transport file to system file and to link data to format library
- If SAS format library, sample code to convert transport file to system file
- Information on Multiple CD Data files, where applicable

Sample "DS" and "DF" README files are contained in Appendix E.

Writing Data and Files onto CD-ROM. Because this is an archiving effort, CD-ROM with data recorded on it should afford read access only. This means that information contained on a CD-ROM generated in this project cannot be modified. This is clearly necessary in an archiving effort, but requires certainty in the information being written during the archiving process. Therefore, the directory structure, README information, and files were initially copied onto media that afford both read and write access. In most cases, the medium used in this intermediate step was the ZIP disk, although hard drives on microcomputers were also used for temporary storage. We modified README files to ensure accuracy, completeness, and uniformity in

[18] When available and consistent between datafiles within a dataset, the project/analysis name from the ARI Research Documentation Database was used. When this name was unavailable or inconsistent, the project names contained in Appendix D were used.

information presentation across ARI projects. Once we were satisfied with what was to be archived, we began copying this information onto CD-ROM using a widely used CD writer software utility.[19] Table 2 presents a summary of the number and types of archived files:

Table 2. Summary of Archived Files

TYPE OF TRANSMITTED FILES	FREQUENCY
Microcomputer Media	
SAS	13
SPSS	96
Flat	11
Other	5
SUBTOTAL	125
Mainframe Media	
SAS	116
SPSS	6
Flat	103
SUBTOTAL	225
Electronic Mail	
SPSS	4
SUBTOTAL	4
TOTAL	354

Archive Quality Assurance Protocols. Over 870 files, including data, format, README, CBK1 and CBK2 files were written during the archiving phase of this project. To ensure accuracy in this archiving effort, we developed three quality assurance protocols. These protocols are contained in Appendix F. Protocol #1 essentially provides a paper trial of all copying activity from ZIP disk or hard drives to CD-ROM. In addition to allowing us to determine which project staff are responsible for the data and its copying onto CD-ROM, it also forces the operator of the CD writing software to be aware of any problems reported by the software.

Quality Assurance Protocol # 2 is the nuts and bolts of our archive quality assurance strategy. In addition to ensuring the accuracy of README file information, this protocol makes sure that all the files that should have been written onto CD were in fact archived. It checks the CD directory structure and CD volume label. It also checks whether codebooks have been archived for the selected Level III data files. This Quality Assurance Protocol was completed by staff other than those who archived the data, so that researchers were not checking their own work. After completion, a senior researcher reviewed all Protocol #2 responses. Based on this review, a few CDs were sent back to the project archivists for revision.

[19] We used the Adaptec Easy CD Creator Version 3.0 to write data onto CD-ROM.

Once CDs passed Protocol #2 and we were thereby assured that their contents were accurate and complete, we affixed labels onto each CD-ROM. These labels contain the ARI project name, dataset number(s), the total number of CDs comprising the project, and the CD number of the current CD. A privacy statement supplied by ARI was also placed on each CD label. CD jewel case inserts were produced containing the ARI logo, the date, the ARI project name, dataset numbers, the total number of CDs for the project, and the CD number of the current CD.

Protocol #3 was then completed to assess the accuracy of the information on CD label and jewel case insert, given the project number internally recorded on the CD volume label. Finally, each data file recorded on the CDs was accessed by senior archivists as a final assurance of the usability of the archived data files.

Summary

All the objectives of this effort have been met. An automated database containing baseline documentation information has been created. Hard copy materials for designated datasets have been organized and archived. All datasets for which component data files were available have been written onto CD-ROM. The following table quantifies the activities and accomplishments of this project:

Table 3. Summary of Project Activities

	Baseline Documentation Level I	Hard Copy Documentation Level II	Codebook Documentation Level III	Data Archived on CD-ROM
# Data Files	542	N/A	14	354
# Datasets	143	37	14	137
# Projects	72	20	12	67
# CD-ROM	N/A	N/A	N/A	93

References

DiFazio, A.S., Young, W. Y., & Driessen, D. P. (1997a). Documentation and archival of selected ARI data bases – final project summary report – phase I (Final Report 97-07). Alexandria, VA: Human Resources Research Organization.

DiFazio, A.S., & Young, W. Y. (1997b). Data base documentation standards for extant datasets (Final Report 97-05). Alexandria, VA: Human Resources Research Organization.

U.S. Army Research Institute (1998). Research, development and acquisition: data documentation and archival policy (Regulation 70-31). Alexandria, VA.

APPENDIX A

Final List of ARI Datasets Documented

Final List of ARI Datasets Documented

Dataset Number	Dataset Name	ARI Assigned Documentation Level	Selected for Documentation in this Effort
3.01	Special Forces Assessment and Selection Database	II	✓
3.02	Special Forces Qualification Course Class Database	II	✓
5.02	Survey of Total Army Military Personnel (1991-1992)	II	✓
6.01	Enlisted Core Dataset-100%	I	✓
6.02	Enlisted Core Dataset-25%	I	✓
8.01	ROTC Commissioned Data Set	I	✓
8.02	United States Military Academy Data Set	I	✓
8.03	ROTC Advanced Camp Data Set	I	✓
8.04	Officer Longitudinal Data Set	I	✓
8.05	Automated Instructional Management System Data Set	I	✓
8.06	Officer Core Data Set	I	✓
8.07	Officer Administrative Data Base (OADB)	I	✓
17.00	Adaptability Screening Profile Faking Data Set (ASP Faking Data Set)	II	✓
20.01	Building the Career Force/Longitudinal Validation	I	✓
20.02	Building the Career Force/Longitudinal Validation I	I	✓
20.03	Building the Career Force/Longitudinal Validation II	I	✓
20.04	Project A/Concurrent Validation I	I	✓
20.05	Project A/Concurrent Validation II	I	✓
21.01	Army Career Transition Survey (ACTS)	II	✓
23.01	Army Family Research Program-Soldiers (AFRP-Soldiers)	III	✓
23.02	Army Family Research Program-Spouses (AFRP-Spouses)	III	✓
27.00	Enlisted Master File	I	✓
28.00	Defense Manpower Data Center Cohort File	I	✓
29.00	MEPCOMS Applicants/Accessions File	I	✓
30.00	ACF Master System File 1981-1986	I	✓
32.00*	Longitudinal Research on Officer Careers (SAS and Flat)	III	
33.00	Proteus 1986, 1987, Merged (86/87)	I	✓
41.00	Army Communications Objectives Measurement System	II	✓
42.00	1990 Army Career Satisfaction Survey	II	✓
43.00	Army Experience Survey	I	✓
45.00	Civilian Leadership (CIVLEAD)	III	✓
49.01	Sample Survey of Military Personnel #1: ENLFEB83	I	✓
49.02	Sample Survey of Military Personnel #2: OFFFEB83	I	✓
49.03	Sample Survey of Military Personnel #3: ENLAUG83	I	✓
49.04	Sample Survey of Military Personnel #4: ENLFEB84	I	✓
49.05	Sample Survey of Military Personnel #5: OFFFEB84	I	✓
49.06	Sample Survey of Military Personnel #6: OFFMAY84	I	✓
49.07	Sample Survey of Military Personnel #7: ENLAUG84	I	✓
49.09	Sample Survey of Military Personnel #9: OFFFEB85	I	✓
49.11	Sample Survey of Military Personnel #11: OFFJUN85	I	✓

* Documentation for this dataset includes an ARI–produced codebook.

** There is no POC-supplied hard copy documentation for this dataset.

Final List of ARI Datasets Documented

Dataset Number	Dataset Name	ARI Assigned Documentation Level	Selected for Documentation in this Effort
49.12	Sample Survey of Military Personnel #12: ENLAUG85	I	✓
49.13	Sample Survey of Military Personnel #13: OFFAUG85	I	✓
49.14	Sample Survey of Military Personnel #14: ENLMAR86	I	✓
49.15	Sample Survey of Military Personnel #15: OFFMAR86	I	✓
49.16	Sample Survey of Military Personnel #16: ENLSEP86	I	✓
49.21	Sample Survey of Military Personnel #21: OFFSPR87	I	✓
49.22	Sample Survey of Military Personnel #22: ENLSUM87	I	✓
49.23	Sample Survey of Military Personnel #23: OFFSUM87	I	✓
49.26	Sample Survey of Military Personnel #26: ENLSPR88	I	✓
49.27	Sample Survey of Military Personnel #27: OFFSPR88	I	✓
49.28	Sample Survey of Military Personnel #28: ENLFALL88	I	✓
49.29	Sample Survey of Military Personnel #29: OFFFALL88	I	✓
49.30	Sample Survey of Military Personnel #30: ENLSPR89	I	✓
49.31	Sample Survey of Military Personnel #31: OFFSPR89	I	✓
49.32	Sample Survey of Military Personnel #32: ENLSPR91	I	✓
49.33	Sample Survey of Military Personnel #33: OFFSPR91	I	✓
49.34	Sample Survey of Military Personnel #34: FALL91	I	✓
49.35	Sample Survey of Military Personnel #35: SPR92	I	✓
49.37	Sample Survey of Military Personnel #37: FALL92	I	✓
49.38	Sample Survey of Military Personnel #38: SPR93	I	✓
49.39	Sample Survey of Military Personnel #39: FALL93	I	✓
50.00	Operation Restore Hope (ORH) in Somalia Spouse Survey	III	✓
53.00**	Gender Integrated Basic Combat Training	II	✓
57.01	Multinational Force and Observers – Sinai: Demographic Dataset (HAWAII_D.SAV)	II	✓
57.02	Multinational Force and Observers – Sinai: Family/Finance Dataset (MERGE3.SAV)	III	✓
57.03	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #1 (G7PLEAD.SAV)	II	✓
57.04	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #2 (G7ELEAD.SAV)	II	✓
57.05	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #3 (G7LLEAD.SAV)	II	✓
57.06	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #4 (G7ALEAD.SAV)	II	✓
57.07	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #5 (G8B4LEAD.SAV)	II	✓
57.08	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #6 (G8PLEAD.SAV)	II	✓
57.09	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #7 (G8ELEAD.SAV)	II	✓
57.10	Multinational Force and Observers – Sinai: Leadership & Climate Dataset #8 (G8LLEAD.SAV)	II	✓
57.12	Multinational Force and Observers – Sinai: Job Knowledge #1	II	✓

* Documentation for this dataset includes an ARI–produced codebook.

** There is no POC-supplied hard copy documentation for this dataset.

Final List of ARI Datasets Documented

Dataset Number	Dataset Name	ARI Assigned Documentation Level	Selected for Documentation in this Effort
57.13**	Multinational Force and Observers – Sinai: Job Knowledge #2	II	✓
57.14	Multinational Force and Observers – Sinai: Job Knowledge #3	II	✓
57.15	Multinational Force and Observers – Sinai: Job Knowledge #4	II	✓
57.16	Multinational Force and Observers – Sinai: Job Knowledge #5	II	✓
57.19	Multinational Force and Observers – Sinai: Other #1 (IDENTITY.SAV)	II	✓
57.20	Multinational Force and Observers – Sinai: Other #2 (CERT1.SAV)	II	✓
59.00	Skill Qualification Test Data Base	I	✓
61.00	Active Army Recruit Database	I	✓
62.00	Bosnia Family Research Database	III	✓
63.00	Officer Standardized Educational Testing Data Base	I	✓
65.00*	Survey of Officer Careers (SOC)	III	
67.00	Special Forces Concurrent Validation	III	✓
68.00**	Special Forces Assessment School (SFAS) Longitudinal Validation	II	✓
71.00	ABLE Coaching Dataset	II	✓
72.00	Fort Jackson Attrition Measures for 1993 Receptees	I	✓
74.01	Non-High School Diploma Graduate (NHSDG) Compensatory Screen Database (FY88-90)	I	✓
74.02	Non-High School Diploma Graduate (NHSDG) Compensatory Screen Database (FY91-92)	I	✓
74.03	NHSDG/HSDG Compensatory Screen Database	I	✓
81.00	Desert Storm Performance Data	III	✓
83.01**	Bosnia Pre-Deployment Soldier Survey	II	✓
83.02	OJE-Soldier During-Deployment Survey	III	
83.03	OJE-Soldier Post-Deployment Survey	III	
87.01	Leadership Requirements: Officers	III	✓
87.02	Leadership Requirements: Non-Commissioned Officers	III	✓
88.00	Montgomery GI Bill 1991-4th quarter 1996	I	✓
89.00	Leader AZIMUTH Check (Ft. Leavenworth Research Unit)	I	✓
90.01	Demographics (BOLDS)	III	
90.02	Cognitive Summer 1994	III	
90.03	Followers, 1996-2	III	
90.04	Followers, 1997-1	III	
91.00	High School Faculty Evaluations of Candidates-West Point USMA	III	
92.00	High School Athletic Activities-West Point USMA	III	
93.00	High School Extracurricular Activities- West Point USMA	III	
94.00	Military Academy Liaison Officer (MALO) Interview Ratings of Candidates- West Point USMA	III	

* Documentation for this dataset includes an ARI–produced codebook.

** There is no POC-supplied hard copy documentation for this dataset.

Final List of ARI Datasets Documented

Dataset Number	Dataset Name	ARI Assigned Documentation Level	Selected for Documentation in this Effort
95.00	Employer Evaluations of Candidates-West Point USMA	III	
96.00	Admissions Data- West Point USMA	III	
97.00	National Survey of Entering Freshmen-West Point USMA	III	
98.00	Class Characteristics Inventory (CCI)- West Point USMA	III	
99.00	Estimated Assessment of Background & Life Experiences (ABLE) and NEO-Personality Inventory Scales Scores- West Point USMA	III	
100.00*	Nelson-Denny Reading Test Scores- West Point USMA	III	
101.00	Cadet Basic Training Leadership Grades and Cadet Performance Report (CPR) Ratings Made by Superiors-West Point USMA	III	
102.00	Term 95-1 Leadership Grades & CPR Ratings Made by Superiors	III	
103.00	Plebe Parent Weekend Duty Positions- West Point USMA	III	
104.00	Term 95-2 Leadership Grades & CPR Ratings	III	
105.00	Cadet Field Training Leadership Grades & CPR	III	
106.00	Term 96-1 Leadership Grades & CPR Ratings by Peers	III	
107.00	Term 96-2 Leadership Grades & CPR Ratings by Peers	III	
108.00	Center for Enhanced Performance Services Utilization	III	
109.00	Leadership Grades and CPR Ratings Made by Superiors During CTLT, DCLT, CBT Trainer, or CFT Trainer Assignments-West Point USMA	III	
110.00	Physical Fitness Performance Measures (APFT and IOCT)- West Point USMA	III	
111.00	Term 97-1 Leadership Grades & CPR Ratings by Peers	III	
500.00	Online Revised Flight Aptitude Selection Test	III	✓
501.00	Alternate Flight Aptitude Selection Test	III	✓
503.00	National Training Center (NTC)	I	✓
510.00	IERW FY 88-90 Data (MAST8890.DAT)	III	✓
511.00	IERW FY 91-92 Data (MAST9192.DAT)	III	✓
512.00	IERW FY 93 Data (MASTER93.DAT)	III	
513.00	IERW FY 94 Data (MASTER94.DAT)	III	
514.00	IERW FY 95 Data (MASTER95.DAT)	III	
519.00	Backward Transfer (Tasks I-XI)	I	✓
520.00	Velocity Vector (Tasks I-VII)	I	✓
521.01	Psychophysical I (X, Y, and Z Axis Combined, All Subjects)	I	✓
521.05	Psychophysical II (All Subjects)	I	✓
528.01	Nuclear, Biological, and Chemical (NBC) Air Warrior Baseline Simulation – Daymaina.xls	I	✓
528.02	Nuclear, Biological, and Chemical (NBC) Air Warrior Baseline Simulation – Nitemana.xls	I	✓
528.03	Nuclear, Biological, and Chemical (NBC) Air Warrior Baseline Simulation – Fire_arc.xls	I	✓

* Documentation for this dataset includes an ARI–produced codebook.

** There is no POC-supplied hard copy documentation for this dataset.

Final List of ARI Datasets Documented

Dataset Number	Dataset Name	ARI Assigned Documentation Level	Selected for Documentation in this Effort
528.04	Nuclear, Biological, and Chemical (NBC) Air Warrior Baseline Simulation – Arc_tix.xls	I	✓
528.07	Nuclear, Biological, and Chemical (NBC) Air Warrior Baseline Simulation – Arc_freq.xls	I	✓
530.00**	A Prototype Army National Guard (ARNG) Armor and Mechanized Infantry Gunnery Training Assessment Database	II	✓

* Documentation for this dataset includes an ARI–produced codebook.
** There is no POC-supplied hard copy documentation for this dataset.

APPENDIX B

ARI Research Dataset Questionnaire

ARI RESEARCH DATASET QUESTIONNAIRE

The purpose of this questionnaire is to collect and store baseline dataset information for all datasets that are submitted to ARI for archive. The information you provide will be added to the on-line ARI Research Documentation Database. Before you begin, be sure to get an ARI Dataset Identification Number from the ARI Science and Technology Officer. The dataset cannot be added to the database until this number is assigned.

Assigned by ARI Science and Technology Officer
ID Number: □

Important

Most of the items in this form are self-explanatory. However, a critical first step for the Point of Contact (POC) is to become familiar with the unit of analysis. Research datasets commonly belong to a group of related datasets. In addition, a dataset may be comprised of many individual data files. There are four typical arrangements by which datasets and data files are organized and stored. Please review the definitions of "databases", "datasets", and "data files" below and consult the Dataset Classification Table on the following page to determine the target "dataset" that should be described in each questionnaire. The table also provides examples of ARI datasets that represent each arrangement and indicates how each arrangement should be described in this questionnaire. A summary of the unit of analysis is also provided on page 3.

- ***Database.*** A database consists of a collection of related, though distinct and often diverse, datasets.
- ***Dataset.*** For purposes of this questionnaire, a collection of conceptually grouped but physically distinct data files is referred to as a dataset. A dataset may have one or more data files.
- ***Data file.*** A data file is a physically distinct collection of data records that contain electronically recorded information generated as a result of data collection activity.

For additional definitions and information, please refer to paragraph 6 of ARI Regulation 70-31. Questions may also be directed to the ARI Science and Technology Officer.

If you need an assigned dataset number or have questions concerning this questionnaire, please contact the ARI Science and Technology Officer at (703) 617-0324.

DATASET CLASSIFICATION TABLE

Parent Database

DEFINITION. For this questionnaire, a database refers to a group of datasets, each containing different types of data, but collectively belonging to, or derived from, the same research project. A parent database is not a physical data file. A separate ARI Research Dataset Questionnaire documents each member of a parent database.

HOW TO COMPLETE. The focus of, or the unit of analysis for, this questionnaire is a dataset. If a dataset belongs to a parent database, complete a separate questionnaire for each dataset. The Parent Database section should be completed in each questionnaire.

Examples:

The Officer Longitudinal Research Data Base (OLRDB) includes: the Core Dataset with personnel data from the annual Officer Master File; the ROTC Dataset with ROTC precommissioning training data; AIMS Dataset with basic and advanced training data; and several other datasets containing different types of data.

Multinational Force and Observers – Sinai (MFO) database consists of the Family/Finance dataset, Cohesion/Leadership datasets, Job Knowledge datasets, and others.

Dataset With Multiple Physical Data Files

all containing a similar set of variables and supporting the same research project or objective

DEFINITION. Multiple data files contain virtually the same number and type of variables, have the same file characteristics, and are stored in the same medium at the same location. The files have different file names, and may differ in number of records, year and/or location of data collection, and the name of computer program used to create the files.

HOW TO COMPLETE. Only one questionnaire is completed for the dataset. However, the Technical Dataset Information section of this questionnaire should be photocopied and completed for **each** physical data file.

Examples:

EMF Dataset consisting of 49 physical files of data extracted from the Enlisted Master File every quarter.

Building the Career Force/Longitudinal Validation I, Batch A MOS dataset consists of 10 SAS system files containing data collected from 20 MOS.

Datasets That Support the Same Project and/or Research Objectives

but differ substantially in contents and/or file characteristics

DEFINITION. Datasets may represent data from a survey that changed significantly in survey format, content, sample group, or phases of a research project (e.g., at enlistment vs. during the second tour).

HOW TO COMPLETE. Complete a separate questionnaire for each dataset even if the responses to many items may be similar (e.g., purpose of the research project, storage medium and location, physical file characteristics, dataset documentation, and extent of current and future use of the data).

Examples:

Bosnia Pre-Deployment, During Deployment, and Post-Deployment Soldier Surveys.

Multiple MFO - Sinai Cohesion/Leadership Datasets differing in the time and location of administration, samples and survey items.

Datasets containing data from the semi-annual Sample Survey of Military Personnel; each survey and the resulting dataset address different issues.

Single, Unitary Dataset

DEFINITION. One physical data file containing all data collected for a specific research project. It may be the end product of merging originally separate files.

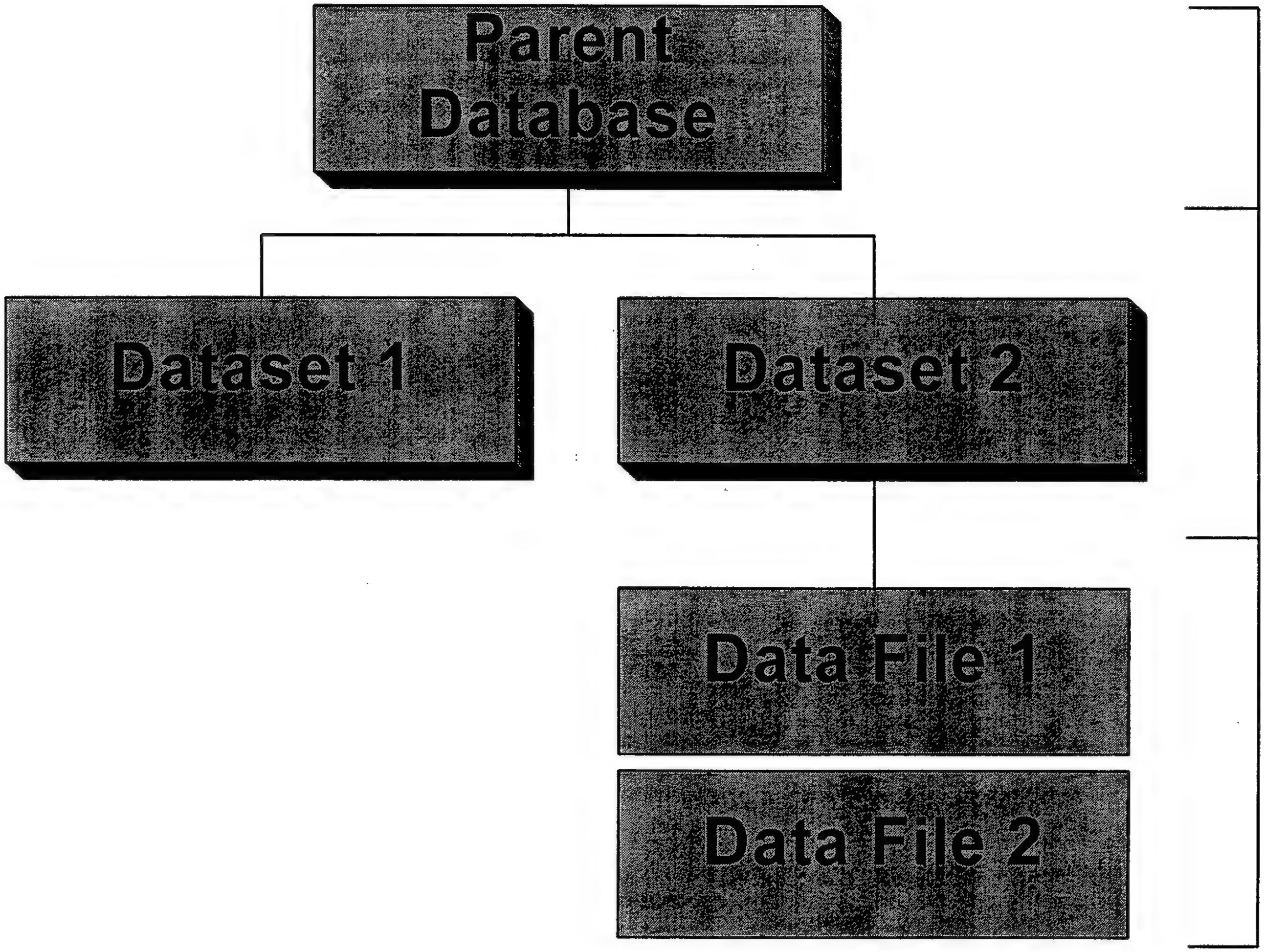
HOW TO COMPLETE. Complete one questionnaire for each unitary dataset.

Examples:

Velocity Vector – one independent dataset containing one data file resulting from the merging of originally separate data files.

ABLE Coaching Dataset – one independent dataset containing one data file.

Survey Unit of Analysis



Parent Database
If a dataset belongs to a parent database, complete the Parent Database section in the survey. Otherwise, omit that section.

Dataset
Complete an entire survey for each dataset.

Data File
A dataset may be composed of multiple physical data files. Complete one entire survey repeating the Techical Dataset Information section for each data file.

ARI RESEARCH DATASET QUESTIONNAIRE

ID Number: □

Point of Contact (POC):

ARI Research Unit:

Dataset Name:

Dataset Acronym:

Brief Dataset Description (Include discussion of the purpose of the research, major areas of investigation, the period covered by the data, and the nature of the data).

General Research Notes:

Available Documentation

The default location for all documents is the Archive Library unless otherwise indicated under Other Location. For documents submitted to the Archive Library, check either electronic or hardcopy.

□ Research Design

□ Electronic

□ Hardcopy

□ Other Location:

□ Sampling Plan and Procedures

□ Electronic

□ Hardcopy

□ Other Location:

□ Data Collection Plan	□ Electronic □ Hardcopy	□ Other Location: ________________
□ Data Collection Instruments	□ Electronic □ Hardcopy	□ Other Location: ________________
□ Dataset Development Process	□ Electronic □ Hardcopy	□ Other Location: ________________

Does the documentation include editing/manipulation specifications? □ Yes □ No

□ Codebook or User's Guide

Does the codebook/user's guide contain the following?

□ Variable list, Proc Contents, or Sysfile information	□ Electronic □ Hardcopy	□ Other Location: ________________
□ Variable descriptions □ Hardcopy	□ Electronic	□ Other Location: ________________
□ Variable code values □ Hardcopy	□ Electronic	□ Other Location: ________________
□ Description of constructed variables □ Hardcopy	□ Electronic	□ Other Location: ________________
□ Variable descriptive statistics □ Hardcopy	□ Electronic	□ Other Location: ________________

□ Record Layout(s)	□ Electronic □ Hardcopy	□ Other Location: ________________
□ Research Report(s)/Publication(s)	□ Electronic □ Hardcopy	□ Other Location: ________________

□ Final Report □ Electronic □ Other Location:

□ Hardcopy ______________________________

□ Other: ________________________ □ Electronic □ Other Location:

□ Hardcopy ______________________________

Comments: ____

Parent Database

Please refer to the definition of a Parent Database located in the introduction to this questionnaire. If the dataset belongs to a Parent Database, complete this section, otherwise proceed to the next Section, Dataset Access.

1. Parent Database Name: ____

2. Parent Database Acronym: ____

3. Can each dataset be used separately for analysis? □ Yes □ No

4. Can the datasets within the database be linked? □ Yes □ No

5. What variable(s) are used to link the datasets (e.g., SSN)? ______________________________

Comments: ____

____

Dataset Access

Please provide information on any restrictions placed on the use of the archived data.

"Individuals requesting data will submit a written memorandum requesting the data through the supervisory chain to the field unit or research unit that has the data in accordance with the Army Research Institute."

Data Sharing Regulation 70-30.

1. Is the access to this dataset or certain variables in the dataset restricted to general users except through the Freedom of Information Act (FOIA) request process (e.g., "close-hold" dataset and/or variables whose release is controlled by a sponsor)?

   □ Yes □ No

2. If yes, does the restriction apply to the entire dataset or certain variables in the dataset?

____

Comments: ____

____

Data Description

This section describes data completeness and accuracy, as well as the research methods and keywords associated with the dataset, population or sample.

1. What methods were used to collect data contained in this dataset? (Check as many as apply.)

□ Interview

□ Performance assessment

□ Cognitive test

□ Attitude assessment

□ Aptitude/achievement tests

□ Observational

□ Survey

Indicate the PT Number (survey approval number from APSO, the U.S. Army Personnel Survey Office): ______________________________

□ Copy of existing operational database

□ Extracted from existing operational database

□ Simulator

□ Other: ____

2. What type(s) of data are contained in this dataset (e.g., personnel, test scores, training performance)?

____

____

3. What years are covered by this dataset?

Fiscal year(s): ____________________ Calendar year(s): ____________________

4. Will this dataset be updated in the future? □ Yes □ No (Go to Question 5.)

a) If yes, what is the planned update schedule? ____________________

b) What is the established update procedure (who does what when)? ____________________

____

5. Does the dataset include unique individual identifiers such as social security numbers?

□ Yes □ No (If no, then go to Question 6.)

a) If yes, are the identifiers encrypted? □ Yes □ No

Note: *Permission must be obtained from the ARI Science and Technology Officer to submit datasets with encrypted identifiers.*

b) Identifier variable name(s) and descriptions: ____________________

____

6. What keywords are relevant to the dataset?

a) Concepts, topic area keywords (e.g., recruitment, re-enlistment, leadership training, family services): ____________________

____

____

b) Measures keywords (e.g., aptitude test): ____________________

____

____

c) Sample groups or population keywords (e.g., commissioned officers, enlisted personnel):

____

____

____

d) Sponsor, related organizational units keywords (e.g.,TRADOC, DCSPER, DCSOPS, FORSCOM, NTC): ____________________

____

____

7. Describe the overall accuracy of the data in this dataset.

____

____

8. Describe the overall completeness of the data in this dataset. If the dataset is designed to include the entire population, do the actual data in the dataset represent the population fully?

____

____

9. Does the dataset have any known bias (due to sampling or data collection procedures)?

____

____

____

10. If the data were collected over time, did the contents of the dataset change over time (e.g., variables, value coding system)?

____

____

____

11. What are the dataset constraints, if any (e.g., size of dataset, too few variables, inability to link with other datasets)?

____

____

12. Representativeness:

a) Does the dataset contain data from a sample(s) of a population or does it describe the entire universe?

□ Sample □ Universe (If a universe, skip to Question 13.)

b) What is the actual, overall representativeness of the sample(s)?

Rating (circle one)

1	Poor	The sample(s)/data are not representative of, or poorly represents, relevant segments of the population
2		
3	Fair	The sample(s)/data approximate some segments of the population, but some specialized areas (e.g., MOS) may not be representative
4		
5	High	The sample(s)/data are representative for every relevant segment of the population

13. Define the universe associated with the data. ____

____

Comments: ____

____

Technical Dataset Information

The purpose of this section is to gather information necessary to locate and read the physical data file(s) associated with the dataset set to be archived. If the dataset contains multiple physical data files, photocopy and repeat this section for each data file. If the dataset contains additional required file(s), such as a format library, repeat this section for each additional file.

1. Dataset management:

   a) ARI Research Unit responsible: ______________________________

   b) Dataset manager: ______________________________

2. What is the archive location of the dataset?

   □ ARI Archive Library

   □ Other/specify: ______________________________

3. Is the dataset compressed? □ Yes □ No

   Note: Special permission must be obtained from the ARI Science and Technology Officer in order to submit data that have been compressed.

   a) If compressed, identify the compression format.

   ____

4. Is the dataset stored in CD ROM format? □ Yes, Number of CDs: __________ □ No

   Note: Special permission must be obtained from the ARI Science and Technology Officer in order to submit data in a format other than CD ROM.

5. What is the dataset name?

   ____

6. Are the data in a structured, system file format dataset (e.g., SAS, SPSS, VSAM, RDBMS) or in raw/flat format (e.g. ASCII)?

   □ System file
   (answer Question 7, skip Question 8)

   □ Raw/flat file
   (skip Question 7, answer Question 8)

7. If the dataset is in a structured, system file format, please provide the following.

Note: Special permission must be obtained from the ARI Science and Technology Officer in order to submit system files that are not in transport or portable format.

a) Are the data in transport or portable format? □ Yes □ No

b) Format (e.g., SAS, SPSS, VSAM, Oracle)? ______________________________

c) Version (if applicable)? ______________________________

d) Table name (if applicable)? ______________________________

e) Name(s) of the program(s) used to create the dataset? ______________________________

8. If the dataset is in raw/flat file format:

Note: Special permission must be obtained from the ARI Science and Technology Officer in order to submit flat files that are not in ASCII format.

a) ASCII format, or other? ______________________________

b) Record format (e.g., fixed length, variable length)? ______________________________

c) Record length (e.g., LRECL)? ______________________________

d) File delimiters (e.g. tab, comma)? ______________________________

9. File size.

a) Number of variables/columns: ______________________________

b) Number of observations/records: ______________________________

10. If this file is a data file and requires additional associated file(s), such as a format library, please check the box below and indicate the name of the additional file(s) in the space provided.

□ Associated File(s) ______________________________

Comments: ______________________________

____

Research Project/Analysis Based On Data From The Dataset

All questions under this section will be completed for each KEY research project or analysis based on data from the dataset. Copy and repeat this section for multiple projects/analyses.

1. Name of project/analysis: ____________________

2. Purpose of research/analysis: ____________________

3. Sponsor: ____________________

4. Instrument clearance approval numbers (e.g., OMB, USAPIC): ____________________

5. ARI Research Unit responsible for research/analysis: ____________________

6. ARI principal investigator: ____________________

7. Begin/end dates of research/analysis: ____________________

8. Description of research/analysis population:

____

____

9. Description of research/analysis sample:

____

____

10. Will similar project/research be conducted in the future?

____

____

Comments: ____________________

____

Publications

If possible complete at least three publications for each related research project or analysis. If the complete information is not available, provide sufficient information, e.g., the document title and/or the first author, to facilitate a search in the Document Archive Retrieval System (DARS).

Title: ____________________

____

Author(s): ____________________

____

Date: ____________ Report #: ____________ Work Unit #: ____________

Title: ____

____

Author(s): ____

____

Date: ______________ Report #: ______________ Work Unit #: ______________

Title: ____

____

Author(s): ____

____

Date: ______________ Report #: ______________ Work Unit #: ______________

Title: ____

____

Author(s): ____

____

Date: ______________ Report #: ______________ Work Unit #: ______________

Title: ____

____

Author(s): ____

____

Date: ______________ Report #: ______________ Work Unit #: ______________

Title: ____

____

Author(s): ____

____

Date: ______________ Report #: ______________ Work Unit #: ______________

APPENDIX C

Sample Level II Documentation Index

ARI Research Dataset Archive

Dataset #65.00
Survey of Officer Careers (SOC)

Index of Hard Copy Documentation

(1) ARI Research Dataset Summary (August 1998)

(2) Data Collection Instrument:

Survey on Officer Careers (1996)

(3) User's Guide/Codebook:

(a) Human Resources Research Organization (1997). Analysis and Reporting Results of 1996 Survey of Officer Careers – Phase I (Final Draft Analysis Plan). Alexandria, VA: Author.

(b) Cowan, C.D. (1996). Documentation of Weighing and Analysis of the Officer's Survey for HumRRO. Alexandria, VA: Author.

(c) U.S. Army Research Institute (1998). Analysis of Differences in Attitudes toward the Army among Racial and Gender Groups (OSC Special Report). Alexandria, VA: Author.

(d) U.S. Army Research Institute (1998). Analysis of Officers' Intentions to Remain with the Army (SOC Special Report). Alexandria, VA: Author.

(e) U.S. Army Research Institute (1998). Officer Attitudes by Commissioning Source (SOC Special Report). Alexandria, VA: Author.

(f) U.S. Army Research Institute (1998). Career Expectations by Type of Branch Assignment (SOC Special Report). Alexandria, VA: Author.

(4) Memorandum:

Memorandum from Jeff Barnes, TO Chuck Cowan, SUBJECT: Survey of Officer Careers Weighing (1996).

(5) User's Guide/Codebook:

Human Resources Research Organization (1996). The 1996 Survey of Officer Careers: Frequencies and Selected Cross Tabulations. Alexandria, VA: Author.

(6) Computer Listings:

(a) SAS PROC CONTENTS Listing of SOC96

(b) SPSS SYSFILE INFO Listing of SOC96C1.SAV

APPENDIX D

List of Documented ARI Projects

List of Documented ARI Projects

Project #	Project Name	# Datasets	# Files	# CD-ROM
3*	Special Forces	2	N/A	N/A
5	Survey of Total Army Military Personnel	1	5	1
6	EPAS (Enlisted Personnel Allocation System)	2	2	5
8	Officer Longitudinal Research Database	7	78	16
17	ASP Faking/Coaching Research	1	2	1
20	Project A/Building the Career Force	5	45	1
21	Army Career Transition Survey	1	1	1
23	Army Family Research Program	2	2	1
27**	Enlisted Master File	1	51	N/A
28**	Cohort Data Base	1	21	N/A
29**	Applicants and Accesions Files	1	104	N/A
30	ACF/MGIB Usage Analysis	1	4	2
32	Longitudinal Research on Officer Careers	1	6	1
33	Proteus	1	3	1
41	Modeling the Individual Enlistment Decision	1	10	1
42	Army Career Satisfaction	1	2	1
43	Army Experience Survey	1	30	1
45	Civilian Leadership	1	1	1
49	Sample Survey of Military Personnel	30	30	1
50	Operation Restore Hope	1	1	1
53*	Gender Integrated Basic Combat Training	1	N/A	N/A
57	Multinational Force & Observers - Sinai	17	17	1
59	Skill Qualification Test	1	11	1
61	Active Army Recruit Database	1	12	2
62	Helping Army Families Cope with Deployment Stress during Operation Joint Endeavor	1	1	1

* These are close-hold datasets.

** ARI will archive these in the future.

*** Two of these datasets are not yet completed.

List of Documented ARI Projects

Project #	Project Name	# Datasets	# Files	# CD-ROM
63	Officer Standardized Education Testing Database	1	1	2
65	Survey of Officers Careers	1	1	1
67	Special Forces Concurrent Validation	1	1	1
68	Special Forces Assessment School Longitudinal Validation	1	1	1
71	ABLE Coaching	1	2	1
72	Best Measures to Predict Attrition at Ft. Jackson	1	2	1
74	Army Compensatory Screen Model	3	8	1
81	Desert Storm Performance Data	1	2	1
83***	Operation Joint Endeavor (OJE)	3	1	1
87	Leadership Requirements Analysis	2	4	1
88	ACF/MGIB Usage Analysis	1	5	1
89	Leader Azimuth Check	1	10	1
90	Baseline Officer Longitudinal Dataset (BOLDS)	4	5	1
91	High School Faculty Evaluation of Candidates	1	1	1
92	High School Athletic Activities	1	1	1
93	High School Extracurricular Activities	1	1	1
94	Military Academy Liaison Officer Interview Ratings of Candidates	1	1	1
95	Employer Evaluations of Candidates	1	1	1
96	Longitudinal Leadership Development Research-West Point Admission Data	1	1	1
97	Longitudinal Leadership Development Research-National Survey of Entering Freshmen	1	1	1
98	Longitudinal Leadership Development Research-Class Characteristics Inventory	1	1	1
99	Longitudinal Leadership Development Research - ABLE & Neo- Personality Inventory Scales	1	1	1

* These are close-hold datasets.

** ARI will archive these in the future.

*** Two of these datasets are not yet completed.

List of Documented ARI Projects

Project #	Project Name	# Datasets	# Files	# CD-ROM
100	Longitudinal Leadership Development Research – Reading Test Scores	1	1	1
101	Longitudinal Leadership Development Research -CPR Ratings	1	1	1
102	Longitudinal Leadership Development Research -Term 95-1 Grades & CPR Ratings	1	1	1
103	Longitudinal Leadership Development Research - Plebe Parent Weekend Duty Positions	1	1	1
104	Longitudinal Leadership Development Research -Term 95-2 Grades 7 CPR Ratings	1	1	1
105	Longitudinal Leadership Development Research - Cadet Field Training & CPR Ratings	1	1	1
106	Longitudinal Leadership Development Research -Term 96-1 Grades & CPR Ratings	1	1	1
107	Longitudinal Leadership Development Research -Term 96-2 Grades & CPR Ratings	1	1	1
108	Longitudinal Leadership Development Research - CEP Services	1	1	1
109	Leadership Grades & Cadet Performance Ratings	1	1	1
110	Physical Fitness Performance Measures	1	1	1
111	Longitudinal Leadership Development Research - Term 97-1 Grades and CPR Ratings	1	1	1
500	Online Revised Flight Aptitude Selection Test	1	1	1
501	Alternate Flight Aptitude Selection Test	1	1	1
503	National Training Center	1	Thousands	5
510	IERW FY 88-90	1	1	1
511	IERW FY 91-92	1	1	1
512	IERW FY 93	1	1	1
513	IERW FY 94	1	1	1
514	IERW FY 95	1	1	1

* These are close-hold datasets.

** ARI will archive these in the future.

*** Two of these datasets are not yet completed.

List of Documented ARI Projects

Project #	Project Name	# Datasets	# Files	# CD-ROM
519	Backward Transfer I-XI	1	11	1
520	AH-64 Velocity Vector	1	1	1
521	Psychophysical	2	2	1
528	Nuclear, Biological & Chemical Air Warrior Baseline Simulation	5	5	1
530	Assessment of the Similar gunnery training strategies through development of a database of gunnery outcome measures	1	1	1

* These are close-hold datasets.

** ARI will archive these in the future.

*** Two of these datasets are not yet completed.

APPENDIX E

Sample README Files

DATASET README INFORMATION
FOR DIRECTORY DS020_02

This directory DS020_02 contains documentation information for ARI dataset # 20.02 "Building the Career Force/Longitudinal Validation I". Documentation and data for each individual data file associated with this dataset are contained in separate subdirectories.

ARI dataset number: 20.02

Dataset Description: The purpose of the "Building the Career Force" component of the Project A was to determine the longitudinal relationship between the new predictors and first-tour performance, to finalize and administer the measures of second-tour job performance, and to examine how selection and classification tests administered before a soldier's first enlistment and measures of performance during their first enlistment predict performance in a second term of enlistment. The Project A longitudinal validation datasets contain predictor data collected from soldiers entering the Army in 1986-87 (at entry into the Army and at the reception stations), as well as criteria measures collected at the appropriate times during soldiers' first and second tour.

The LVI datasets contain data collected (in July 88 through February 89) from the 86-87 cohort on a full array of first-tour job performance measures when they had between 18 and 24 months of service.

Project/analysis name: Project A/Building the Career Force

Archive date: April 1997

Number of data files associated with the dataset: 12

Names and contents of files and subdirectories within directory DS020_02:

1. DF001: Contains SAS transport file, codebook file and documentation for M7A11BV6.XPT (SAS.M7A11BV6)

2. DF002: Contains SAS transport file, codebook file and documentation for M7A13BV6.XPT (SAS.M7A13BV6)

3. DF003: Contains SAS transport file, codebook file and documentation for M7A19EV6.XPT (SAS.M7A19EV6)

4. DF004: Contains SAS transport file, codebook file and documentation for M7A19KV6.XPT (SAS.M7A19KV6)

5. DF005: Contains SAS transport file, codebook file and documentation for M7A31CV6.XPT (SAS.M7A31CV6)

6. DF006: Contains SAS transport file, codebook file and documentation for M7A63BV6.XPT (SAS.M7A63BV6)

7. DF007: Contains SAS transport file, codebook file and documentation for M7A71LV6.XPT (SAS.M7A71LV6)

8. DF008: Contains SAS transport file, codebook file and documentation for M7A88MV6.XPT (SAS.M7A88MV6)

9. DF009: Contains SAS transport file, codebook file and documentation for M7A91AV6.XPT (SAS.M7A91AV6)

10. DF010: Contains SAS transport file, codebook file and documentation for M7A95BV6.XPT (SAS.M7A95BV6)

11. DF011: Contains SAS transport file, codebook file and documentation for M7AMOZV1.XPT (SAS.M7AMOZV1)

12. DF012: Contains SAS transport format library (SASFMTS.XPT) for all 11 data files

13. CBK2.DOC: Contains Codebook Part II common to DF001-DF010

14. README.TXT: Dataset information for directory DS020_02

DATA FILE README INFORMATION
FOR SUBDIRECTORY DF001 IN DIRECTORY DS020_02

This subdirectory DF001 within directory DS020_02 contains documentation information for the first data file associated with ARI dataset #20.02 "Building the Career Force/Longitudinal Validation I" named M7A11BV6.XPT.

ARI dataset number: 20.02

Project/analysis name: Project A/Building the Career Force

Archive date: April 1997

Data file name: M7A11BV6.XPT

Contents of subdirectory DF001:

1. SAS Transport File: M7A11BV6.XPT
   Original file name: SAS.M7A11BV6

2. Codebook: CBK1.DOC
   CBK2.DOC (stored in directory DS020_02)

3. Data File Information: README.TXT

To create a SAS system file on the C drive using the SAS transport file on this CD-ROM, where the E drive is the CD reader drive:

```
libname trans xport 'e:\ds020_02\df001\m7a11bv6.xpt';
libname out1 'c:\';
proc copy in=trans out=out1;
proc contents;
run;
```

To create a SAS format library on the C drive using the SAS transport format library on this CD-ROM, where the E drive is the CD reader drive:

```
filename tranfile  'e:\ds020_02\df012\sasfmts.xpt';
libname out1 'c:\';
proc cimport library=out1 infile=tranfile;
run;
```

To link the SAS system file and the SAS format library and generate frequency distributions of selected variables:

```
libname out1 'c:\';
libname library 'c:\';
proc freq data=out1.m7a11bv6(keep=m7race m7sex);
run;
```

APPENDIX F

Archive Quality Assurance Protocols

DATADOC
ARCHIVE QA PROTOCOL #1

ZIP DISK → CD ROM COPIES

ZIPDISK #: ______

ZIPDISK SOURCE PERSON:

Maggie_______

Name of PROJ Directory Copied: __________________

Winnie _________

Name(s) of DS Directories Copied:

____

____

____

____

____

____

____

____

____

____

____

____

PROJECT NUMBER: _____ *(This is the common integer portion of the dataset numbers being copied.)*

DATE COPIED: ____/____/____

NAME OF PERSON WHO PERFORMED COPY _____________________

COPY UTILITY ENDED SAYING COMPLETED WITH NO ERRORS?
Yes_______ No______

DOES THE JEWEL CASE HAVE A YELLOW STICKY WITH THE PROJECT NUMBER ON IT? Yes_______ No______

CD # (project #): ____________

DATADOC

ARCHIVE QA PROTOCOL #2

DIRECTORY STRUCTURE AND README FILE ACCURACY

NOTE: MAKE AS MANY COPIES OF SECTIONS B-C OF THIS QA PROTOCOL AS THERE ARE DATASET NUMBERS BELOW. FILL IN THE APPROPRIATE DATASET NUMBER IN EACH SECTION B.

A. CD ROM ROOT DIRECTORY

1. List all the dataset numbers for this project from the ARI Datasets Status List.

__________________	__________________
__________________	__________________
__________________	__________________
__________________	__________________
__________________	__________________
__________________	__________________
__________________	__________________
__________________	__________________

2. Using Window's Explorer, does the CD root directory contain subdirectories for all the datasets for this project (i.e., is there a DS_ directory for each dataset listed in A1)?
   Yes _______ No _______

2a) Using Window's Explorer, does the CD ROM have an internally recorded "PROJECT_#" and is it correct?
   Yes _______ No _______

B. DS DIRECTORY

DATASET ID NUMBER ____________________

1. Using Window's Explorer, does the DS directory have a README.TXT file?
   Yes _______ No _______

2. List the file names that the *Technical Dataset Information* section of the ACCESS QUESTIONNAIRE (Tables 1 and 2) indicates that there should be for this dataset (please include all secondary associated files, such as format libraries):

1) ________________
2) ________________
3) ________________
4) ________________
5) ________________
6) ________________
7) ________________
8) ________________
9) ________________
10) ________________
11) ________________
12) ________________
13) ________________
14) ________________
15) ________________
16) ________________
17) ________________
18) ________________
19) ________________
20) ________________
21) ________________
22) ________________
23) ________________
24) ________________
25) ________________
26) ________________
27) ________________
28) ________________
29) ________________
30) ________________
31) ________________
32) ________________

3. Using Window's Explorer, are there as many DF directories on the CD ROM as there are datafile names listed in B2 above?
   Yes _______ No _______

NOTE: MAKE AS MANY COPIES OF SECTIONS D-E AS THERE ARE DFs LISTED ABOVE. FILL IN THE APPROPRIATE DF NUMBER (001, 002, etc) IN EACH SECTION D.

C. README.TXT FILE IN THE DS DIRECTORY

TITLE LINE

4. Does the title read exactly "DATASET README INFORMATION FOR DIRECTORY DS…_..."? Yes _______ No _______

5. Does the title list the correct DS directory name (i.e., does the directory name contain the Dataset Number you wrote in the beginning of section B above, with a DS before it and an underscore instead of a decimal point)?
Yes _______ No _______

>>>>>>> BLURB UNDER TITLE LINE

5a. Does the blurb read exactly as follows:

This directory DS…_.. contains documentation information for ARI dataset # ___.__ "(this stuff in quotes may be left out if the descriptive dataset name below it is unavailable)". Documentation and data for each individual data file associated with this dataset are contained in separate subdirectories.
Yes _______ No _______

6. Is the DS directory name in the first line of the blurb the same as the directory name in the Title line?
Yes _______ No _______

7. Is the ARI dataset number in the first sentence of the blurb the same dataset number as that embedded in the DS directory name?
Yes _______ No _______

8. Is the name of the ARI dataset in quotes the same as the Dataset Name on the top line of the first page of the ACCESS QUESTIONNAIRE?
Yes _______ No _______

>>>>>>>>> INFORMATION BELOW THE BLURB

8a. Do the subsection headings appear exactly as follows:

ARI dataset number:
Dataset description:
Project/analysis name:
Archive date:
Number of data files associated with the dataset:
Names and contents of files and subdirectories within directory ________:
Yes _______ No _______

9. Is the dataset number listed beside "ARI Dataset Number" correct, (i.e., is it the same as that embedded in the DS directory name, title line, and in the blurb)?
Yes _______ No _______

10. Is the name provided in "Dataset Description" correct (i.e., the same as in the first page of the ACCESS QUESTIONNAIRE under "Dataset Description")?
Yes _______ No _______

11. Is the "Project/analysis name" listed correct, (i.e., if the "Project/analysis name is filled out in the ACCESS QUESTIONNAIRE, then it should be exactly that. If it is missing, then it should be exactly the name listed on our 5/28/98 list DATADOC PROJECT NAMES)?
Yes _______ No _______

12. Is the number listed for "Number of data files associated with the dataset" the same as item # B2 above? Yes _______ No _______

12a Is the number listed for "Number of data files associated with the dataset" the same as that reported for the dataset in the ARI DATASET STATUS LIST?
Yes _______ No _______

13. Is the DS directory name listed in "Names and contents of files and subdirectories within directory ______" the correct one, (i.e., is it the actual directory name)?
Yes _______ No _______

13a. None of these subheadings should be blank after the semicolon. If there is no information to fill in one or more of these subheadings, does it state explicitly "Information not available"?
Yes _______ No _______ Does not apply _____

>>>>>> LIST OF FILES UNDER "NAMES AND CONTENTS OF FILES..."

13b. Are there as many DF subdirectories listed here in the README file as there are files listed in response to item B2 above?
Yes _______ No _______

14a. Using Window's Explorer to display the contents of each DF subdirectory, list the data (not README) file names associated with each DF.

DF001______________________________

DF002______________________________

DF003______________________________

DF004______________________________

DF005______________________________

DF006______________________________

DF007______________________________

DF008______________________________

DF009______________________________

DF010______________________________

DF011______________________________

DF012______________________________

DF013______________________________

DF014______________________________

DF015______________________________

DF016______________________________

DF017______________________________

DF018______________________________

DF019______________________________

DF020______________________________

DF021______________________________

DF022______________________________

DF023______________________________

DF024______________________________

DF025______________________________

DF026______________________________

DF027______________________________

DF028______________________________

DF029______________________________

DF030______________________________

DF031______________________________

DF032______________________________

14b. Are the names of the files listed in C14a the same as those listed here in the README file? Yes _______ No _______

14c. Do each of these file names appear in the *Technical Dataset Information* section of the ACCESS QUESTIONNAIRE (Tables 1 & 2)?
Yes _______ No _______

15. In the README file, are DF files listed first, followed by the README.TXT? Yes _______ No _______

16. a) If the *Technical Dataset Information* section of the ACCESS QUESTIONNAIRE (Tables 1 & 2) indicates that the DF file is a SAS or SPSS file, is it described here in the README.TXT as a transport file?
Yes _______ No _______ Does Not Apply ______

b) If yes to C16a, then is the file simply called a SAS transport file rather than SAS transport system file?
Yes _______ No _______ Does Not Apply ______

c) If the *Technical Dataset Information* section of the ACCESS QUESTIONNAIRE indicates that the DF file is a flat file, is the word 'transport' left out of README.TXT?
Yes _______ No _______ Does not apply _____

d) If yes to C16c, is the file called a flat file (rather than ASCII or raw file)?
Yes _______ No _______ Does not apply _____

17. Using the ARI DATASET STATUS LIST, is the dataset a level III dataset that was chosen for codebook development?
Yes _______ No _______ (if no, go to question C18)

18. Are the words "and documentation" (rather than readme files, codebook files, etc) part of the brief DF description?
Yes _______ No _______

19. a) Is there a DF with a format library(s) listed (i.e., is there a file named SASFMTS.XPT?
Yes _______ No _______ (if no, go to question C20)

b) If yes to C19a, is the format library listed in the README file as a transport format library?
Yes _______ No _______ Does Not Apply ______

c) Is this format library(s) listed in the *Technical Dataset Information* section of the ACCESS QUESTIONNAIRE (Tables 1 & 2)?

Yes _______ No _______ Does Not Apply ______

d) Does the README.TXT file indicate which datafiles the format library applies to?

Yes _______ No _______ Does Not Apply ______

e) Does the information in the README.TXT file about which datafiles the format library applies to agree with the information in the ACCESS QUESTIONNAIRE (Tables 1 & 2)?

Yes _______ No _______ Does Not Apply ______

20. a) Is the README.TXT file listed as a file in DS directory README file?

Yes _______ No _______

b) Is the directory name in the README reference correct?

Yes _______ No _______

D. DF___ (fill in #) SUBDIRECTORY

(NOTE: Make as many copies of Sections D and E as there are files reported in item C14A above)

1. Using Window's Explorer, does the DF subdirectory contain the datafile listed in item C14a above for this DF?
   Yes _______ No _______

2. Using Window's Explorer, if item C17 above indicates that the DF should have a .DOC files, are they contained in the DF subdirectory?
   Yes _______ No _______ Does not apply _____

3. Using Window's Explorer, is there a README.TXT file in the DF subdirectory?
   Yes _______ No _______

E. README.TXT FILE IN THE DF DIRECTORY

TITLE LINE

4. Does the first line say "DATA FILE (rather than dataset) README INFORMATION"? Yes _______ No _______

5. Does the second title line list the correct DS directory name (the one listed in the corresponding Section B and the DS directory that the DF is actually stored in)?
   Yes _______ No _______

6. Does the second title line list the correct DF subdirectory name (the one listed in the corresponding Section D)?
   Yes _______ No _______

BLURB UNDER TITLE LINE

7. Does the blurb read exactly as follows:

   This subdirectory DF--- within directory DS---_-- contains documentation information for the ___ (first, second, etc) data file associated with ARI dataset #---.—"(this stuff in quotes may be left out if the descriptive dataset name in the associated DS directory is unavailable)" named _____.___.
   Yes _______ No _______

8. Is the DF subdirectory number (e.g., 001, 002) in the first line the same as the subdirectory name and the DF subdirectory number in the title?
   Yes _______ No _______

9. Is the DS directory name in the first line the same as the directory name in the Title line?
   Yes _______ No _______

10. Is the ARI dataset number (e.g., #91.00) the same dataset number as that embedded in the DS directory name?
    Yes _______ No _______

11. Is the name of the ARI dataset correct for that dataset number, (i.e., is it the same name as that provided on item C8 above)?
    Yes _______ No _______

12. Is the file name the right one, as listed in item #C14a above?
    Yes _______ No _______

INFORMATION BELOW THE BLURB

13. Do the subsection headings appear exactly as follows:

    ARI dataset number:
    Project/analysis name:
    Archive date:
    Data file name:
    Contents of subdirectory DF___:

    Yes _______ No _______

14. Is the 'ARI DATASET NUMBER:' correct (i.e., is it the same as that reported in the blurb above)?
    Yes _______ No _______

15. Is the 'PROJECT/ANALYSIS NAME:' the same as that in the associated DS README?
    Yes _______ No _______

16. Is the 'DATAFILE NAME:' the correct one (same as that in C14a and in the blurb above)?
    Yes _______ No _______

CONTENTS OF DF SUBDIRECTORY

18. Using Windows Explorer, are the files listed in the DF README.TXT the same as those that are actually on the CD?
Yes _______ No _______

19. a) If a SAS transport file is listed, is the filename suffix '.XPT'?
Yes _______ No _______ Does Not Apply ______

b) If the filename has a suffix '.XPT', is it called a ' SAS TRANSPORT' file?
Yes _______ No _______ Does Not Apply ______

c) If an SPSS transport/portable file is listed, is the filename suffix '.POR'?
Yes _______ No _______ Does not apply _______

d) If the filename has a suffix '.POR', is it called a 'SPSS TRANSPORT or PORTABLE' file?
Yes _______ No _______ Does not apply _______

e) If the file is an ASCII file, is it referred to as a FLAT FILE in the DF README?
Yes _______ No _______ Does not apply _______

20. If the file is data file (not a format library), is the original file name given?
Yes _______ No _______ Does Not Apply ______

22. a) If the answer to C17 is "yes", are there CBK1.DOC and CBK2.DOC files listed in the README.TXT file as contents of the DF directory?
Yes _______ No _______ Does not apply _______

b) If the answer to C17 is "yes", using Windows Explorer, are there CBK1.DOC and CBK2.DOC files actually recorded in the DF directory?
Yes _______ No _______ Does not apply _______

(23. – Blank Item)

24. a) Is the README.TXT file always the last file mentioned?
Yes _______ No _______

b) Is the README.TXT file provided following the words 'DATA FILE INFORMATION:'

Yes _______ No _______ Does not apply _______

INFORMATION FOLLOWING CONTENTS OF DF DIRECTORY

25. a) If the file is not a SAS transport file, is the rest of the README file blank?

Yes _______ No _______ Does not apply _______

b) If the file is a SAS data file, is there code provided to create a system file from the transport file?

Yes _______ No _______ Does not apply _______

c) If the file is a transport SAS format library, is there code to create a SAS format library from the transport library?

Yes _______ No _______ Does not apply _______

d) If the file is a SAS data file, is there a SAS format library elsewhere on the CD?

Yes _______ No _______ Does not apply _______

e) If 1) yes to D25d, then is there code provided to a) create a format library from the transport library, and b) link the data file to the format library or 2) no to D25d, then is that code omitted?

Yes _______ No _______ Does not apply _______

DATADOC
ARCHIVE QA PROTOCOL #3

CD VOLUME LABEL AND JEWEL CASE INSERT ACCURACY

1. Using Windows Explorer, does the project number on the internally recorded volume label match the integer portion of the DS and DF directory names?
   Yes_____ No_____

2. Is there a jewel case insert for this project?
   Yes_____ No_____

3. Is the project number and name on the jewel case insert correct for that CD?
   Yes_____ No_____